【05】第4章 | 收集高质量的数据

在「序篇」中我们提到过，高效的数据收集方式得到的数据，应该是标准化和结构化的，这两点对于数据的处理至关重要。不过，如果我们的目标是收集高质量的数据，那么高质量的定义，显然要更宽泛些。事实上，标准化和结构化的数据，只是方便我们进一步地使用数据，这只是高质量数据的特征之一。而要完整地定义高质量的数据，则需要从三个方便入手：

- **真实性**确保了数据**有用**：显然，没有人希望收集到的数据是虚假的。失去了数据真实的前提，也就没有了分析数据和使用数据的价值；

- **结构化和标准化**使得数据**方便用**：在收集到数据后，往往需要进一步的加工、处理、分析和应用。结构化和标准化的数据，有助于大大提升我们在后续工作中的效率；

- **和人关联**进一步让数据**好用**：数据是死的，人是活的，尤其对于当代职场人来说，许多时候收集数据是为了更好地推进各类运营工作。那么，真正好用的数据，能够让我们透视到数据背后的人，从而描画出用户的具体画像，方便我们精准推送、触达和召回用户；

那么，如何让金数据帮助我们收集到高质量的数据呢？

保障数据的真实

数据的真实体现在填写人和填写内容的真实。除此之外，针对投票、抽奖类的表单，哪怕填写人和内容都是真实的，也需要杜绝一个人反复刷票。

1. 填写人真实

开启手机号短信验证

就像注册需要填写手机验证码一样，金数据支持在填写表单中的手机字段时，必须输入短信验证码后才能提交表单。这样，就能够最大程度地保证，用户留下的手机号是真实有效的。由于每次发送短信都是有成本的，建议同时开启「获取验证码前需填写图形验证码」的选项，避免短信验证码的滥发。

限定字段白名单

在运营工作中，我们经常只想收集核心忠诚用户的反馈，并给予一定的奖励。这种情况下，我们的调研表单只希望针对特定的群体，而不希望任何人都能填写。或者，在收集公司文化衫这样的表单时，我们希望填写对象仅限于企业内部的员工。这时候就可以使用金数据的字段白名单功能。

针对**单行文字**、**姓名**和**手机**这三种类型的字段，你可以限定用户填写的值，必须和你预设的白名单相符，否则就不能提交表单。这样，你可以提前把企业员工的姓名或忠诚用户的

手机号预设在表单中，只有这些特邀用户才能够提交表单。

2. 填写内容真实

想要让用户提交的表单内容，都是发自内心的答案，而不是胡乱选择的？我们当然没有办法 100% 地做到这一点，不过，通过这一些选项，可以最大程度地限制或排除掉不真实的内容。

随机选项顺序

在一些调研类的表单中，经常会有多个询问用户满意程度的问题，由于答案的排列一般都是「非常满意、满意、不满意、非常不满意」这样的顺序，导致许多人在几道题目之后，为了图省事，就开始无脑地勾选某个固定位置的答案，而不再真的认真思考作答。

为了不让用户在选项间形成选择惯性，针对**单项选择**、**多项选择**、**图片单选**和**图片单选**这四种类型的字段，可以开启「选项展示顺序随机」，这样，用户每次打开表单后，看到的可选项的排列顺序就不再是固定的了。

不能和已有数据重复

假设你是一位 macOS app 开发者，近来你决定将用户从 App Store 迁移到自有的商店渠道。针对所有已经在 App Store 付费购买的用户，只需要通过表单提交自己的邮箱和购买的订单号，你就可以给他们发送一个新的自有商店版本的授权激活码。

不过，你也遇到了一个小小的问题：有一个用户把自己的订单号公布在了网上，导致许多贪小便宜的人，都用这个订单号来申请激活码了。这时候该怎么办？

针对**单行文字**、**邮箱**、**网址**、**姓名**、**电话**和**手机**这六个类型的字段，你可以开启「不能和已有数据重复」，从而避免用户提交重复的内容。

设置为「不能和已有数据重复」，当已经有人提交过相同内容时，后续的人将不能提交表单

事实上，这个功能提供了一个重要的特性：即只接受首次输入某个值的用户。想一想，它希望还有很多好玩的使用场景，例如：

- 允许用户自选靓号：首个提交该号码的人，拥有该靓号的使用权，之后的人无法选中同样的靓号；
- 竞答活动只有一个胜出者：猜灯迷活动，一下子就能知道谁是第一个答对的人；猜数字大小，只有一个获奖者，因为正确答案在表单中都是唯一的；

在设置问题时交叉验证

在评估表单内容有效性的时候，问题间的交叉验证是一个常用的手段。例如，你可以设置两个问题：

1. 你什么时候第一次使用我们的产品？
2. 你什么时候第一次听说我们的产品？

如果在作答问题 1 时，回答第一次使用是五年前，而问题 2 第一次听说则是近一年内，则显然是在瞎答了。这整个提交内容的可信度就不高了。因此，你可以通过一些题目的设置，如果用户的作答出现明显的自相矛盾，在后期处理数据时，可以将其剔除在外。

3. 防止刷票

在投票、调研、抽奖类的表单中，为了防止有人反复填表和恶意刷票，可以采取如下措施：

限制填写人的身份与填写频率

在「设置 - 数据收集」中，可以限制填写者的权限。例如，当允许所有人可填时，你可以限制每人限填一次。

- 每个 IP 限填一次：同一个 IP 地址，只能填写一次表单，这个限制较严格，但相应的误杀概率也较大。例如网吧、办公场所、学校等地方，往往都是同一个 IP 地址，办公室中一个人填写了，别人就没法再填了。

- 每台电脑/手机限填一次：这个限制是基于电脑本地的 Cookies 的，一旦填写过后，会在 Cookies 中保存已经填过的状态。不过，对于具有一定计算机知识的人来说，可能会绕过这个限制。因为 Cookies 状态是本地的，只需要清空本地浏览器的 Cookie 状态，或进入隐身模式等，就可以使该限制失效。

可以选择较宽松的按电话/手机限填，或较严格的每个 IP 限

填一次，但后者有可能造成误伤

由于这样的限制要不误杀太严重，要不太容易绕过，如果真的想限制每个人只能填写一次，可以试试「仅金数据用户可填」。启用这一选项，会要求填表人必须注册登录金数据的账号体系，从而进一步增加了一个人反复填表的成本。

开启提交验证码

为了防止机器自动刷票，你还可以在「设置 - 数据收集 - 填写设置」中开启「提交前需填写验证码」的选项，在每次提交表单时，都需要人工点选验证码正确后，才能够提交表单。

开启提交前填写验证码，每次必须选对验证码后，才能提交表单

金数据的验证码使用的是图形选择题，每次需要根据提示内容，点选正确的图标。当然，所有的这些限制措施，也会阻挠正常用户的表单填写体验。如果你同时开启了手机号验证、限金数据登录用户填写、提交验证码，那么势必会大大影响表单的填写完成率。因此，你自己也需要权衡，这份表单的目标，是为了精准地收集高质量的数据，宁可错杀一百也在所不惜；还是不需要那么严格，宁愿收集一些垃圾数据，自己在后期去处理剔除。

规范数据的结构化和标准化

数据的结构化和标准化，有助于提升我们进一步处理和使用

数据的效率。告别传统的微信、邮件收集方式，转而使用表单工具，已经做到了基本的结构化。那么，要想做到标准化，还需要什么呢?

1. 选用正确的字段类型

同样的问题，其实可以选择不同的字段类型。例如，在收集员工的文化衫尺码时，你既可以设置成「单行文字」题让大家自由输入，也可以设置成「单项选择」题从几个选项中选择。不过，前者收集回来的数据可就五花八门了，同样要表达 XL 码，大家的填写方式可能有「大码、XL码、XL、xl、Ｘ Ｌ」等等。

因此，在选用字段类型时，谨记这几个原则：

- **相比单行文字题，尽可能使用单项选择题**：如果选项是可以穷尽的，那么一定要设置成单项选择题。如果真的有极少部分的特殊情况，那也在单项选择题中，设置一个「其它」的选项，并让用户填写补充说明。

  使用单项选择题时，如果有选项不能覆盖的情况，可以让用户在其它选项中补充说明

- **让用户打分选择「评分」题**：如果你想向用户评分，可以选用「评分」字段，它支持三档满分制：3 分/5 分/10 分，并且有图形化的展示界面可以直接点选。如果选用「数字」字段，用户需要从键盘输入，还可能会输入小数。

- **输入身份证号选择「单行文字」题**：金数据的「数字」题，最多只能输入 15 位数字。如果你想收集用户的身份证号信息，可以选用「单行文字」题，它还可以开启「身份证号验证」。不过，它只是校验身份证号格式是否符合规则，并不能判断身份证号的真伪。

  身份证号验证可以校验身份证是否合法，但不能校验真伪

- **使用联系信息字段**：当你想要收集用户的姓名、手机、邮箱、地址、电话等信息时，切记使用金数据提供的「联系信息」字段。尤其是手机、邮箱字段，如果你使用「单行文字」字段来收集，虽然同样可以达到效果，但是就无法调用金数据的收发短信/邮件功能了。

  只有使用联系信息中的字段，填写的手机、邮箱才会被识别成为可以群发短信/邮件的联系人

2. 设置字段的校验规则

字段的校验规则，可以让你规定每个字段的合法输入值。例如，在收集用户的建议反馈时，我们可以规定最小的填写字数，避免用户随便写两个字，提交无效的信息。

在收集用户的主观反馈时，可以要求必须最少输入 10 个字，减少低质量的内容提交

具体来说，以下字段支持添加校验规则：

- **数字**：可以限制填写的最小值/最大值，如面向青少年的问卷，可以规定年龄必须在 12–20 岁之间；
- **日期**：可以限制可选择日期的起止时间，如选择参会时间，可以限定在会议有效期内；
- **单行文字** 和 **多行文字**：可以限制填写内容的长度，最少/最多填几个字，如用户建议反馈最少 10 个字；
- **多项选择** 和 **图片多选**：可以限制最少/最多选几项，如了解用户最常用的社交网络，最多只能选 3 项；

3. 引导用户按照统一的标准输入

在选用正确类型的字段，并设置好校验规则之后，其实我们已经完成了**客观上**的标准化。然而，标准化的工作并没有完成，为了最终的确的数据标准化，我们还必须引导用户做到**主观上**的标准化。

比如说，对服务满意度打分，当我们选用「评分」字段，且设置为 10 分满分制的时候，这个题目在客观上已经做到了标准化。然而，不同的人打分的主观标准不同，有的人觉得不错统统打 10 分，有的人只打 5 分。这样，最后的数据就容易有误差。

同样的道理，衣服的尺码，虽然只分成 S/M/L/XL 几档，然而，不同的品牌，尺码的大小标准可能是不统一的，身高 178 究竟是选 L 码，还是 XL 码呢？

这就需要我们在字段的描述中，做出更多引导性的叙述，促使用户按照统一的标准输入。例如在文化衫收集的案例中，我们详细列出了尺码和身高的对应关系，并注明是男女均码。同样的道理，在打分时，你也可以引导告知用户「感觉一般请打 5 分；如果你强烈想把我们的产品推荐给朋友，请打 10 分；如果你再也不会使用我们的产品，且会阻止身边的人使用，请打 1 分；」

在字段的描述中加以明确的说明，尽可能地引导不同人按统一的标准填写

让数据和人关联

做到了前两点，收集到的数据在单次看来，已经质量不错了。然而，如果我们最终的目标，并不是收集一次孤立的数据，而是希望能够积累使用，或洞察数据背后的人，那么让数据和人关联这件事情，就显得尤为重要了。想想以下场景，其实都需要把数据和人关联起来：

- 针对高净值用户的深度营销：「英语棒」试听课过后，那些满意度较高的用户，更有可能购买正式课程，这些人群需要重点进行深度营销；

- 针对流失用户的召回：如果一个用户去年频繁地填写活动报名表，积极参与线下活动，今年却不怎么来了，我们就需要针对这部分流失用户做召回；

- 免去用户重复填写的麻烦：今年收集过文化衫的尺码

了，明年定做新的文化衫时，那么填过的员工就可以不必再填了；

- 分析用户的属性画像：使用金数据搭建了一个小型电商平台，我们肯定想进一步了解购买人群的地理分布、性别比例等属性；

1. 开启微信增强

在「设置 - 微信增强」中，可以开启收集填写者的微信个人信息。开启后，可以收集到微信用户的昵称、头像、性别、国家、省市以及 OpenID。这些信息都是通过微信公众平台的开放接口获取到的，并且，OpenID 并不同于微信用户的手机号或微信号，而是一个没有规则的字符串标识，它对应每个公众号都是不一样的，也没有办法用它来添加某人的微信好友。

开启微信增强中的收集个人信息功能，可以获取到用户的微信名、头像、昵称、省市等信息

这里需要注意的是，金数据提供了两个配置：

- 收集主体：如果你有自己的认证微信公众号，可以关联绑定你自己的公众号；如果没有，则使用金数据的官方微信认证号。区别在于前面提到的，对应同一个用户，不同公众号收集到的 OpenID 也是不同的。如果你自己还有一些业务会需要和表单填写人的微信 OpenID 相关联，那么使用自己的认证微信公众号，会更方便后续的

用户身份匹配工作。

- 授权方式：可以选择静默授权或者点击授权。区别在于，前者对于填表人是没有感知的，只收集用户的微信 OpenID，而不获取昵称、头像、性别、国家、省市等信息；后者则可以拿到所有的信息，但需要用户主动点击授权才可以（但如果用户已经关注了你选用的收集主体，则不需要点击确认了）。

如果你选择的是「填表前需点击确认」的授权方式，最终会在收集到的数据中，看到新增的额外几列，展示了用户的个人信息。

开启收集微信用户信息后，会在最终收集到的数据里，显示额外的微信身份信息

2. 和自有系统中的用户身份关联

如果你有一个自己的 app，在 app 内让用户填写新功能反馈时，我们往往想知道是哪位用户填写的。然而，由于用户已经在你的 app 里了，他并不清楚打开的表单页面是金数据的，这时候还要他再输入一遍自己的用户名等其它信息，他会觉得特别奇怪，我明明登录在你的 app 里啊。

金数据也可以解决这一点，并且解决方案我们在「第三章」中已经提到过了：扩展属性。那时候我们用它来统计不同渠道的表单回收率，例如在后面加上 weixin 后，把这个链接在微信环境中传播，收集回来的数据就会有一列显示 weixin。

在扩展属性中，不一定添加的属性是渠道，也可以将用户的唯一 ID 标识作为参数填入

现在，你只需要让研发人员帮一个小忙，当用户在 app 内点击打开金数据的表单时，在链接后面拼接上自有系统的用户唯一 ID 标识，最终就能对应上这条数据是哪位用户提交的了。

jinshuju.net/f/{表单id}?x_field_1={让研发人员用代码拼接上自有系统的用户唯一 ID 标识}

不过，需要注意的是，通过这种方式并不是验密的，也就是说任何人只要拿到链接和用户 ID，都可以打开这个页面。因此，只能用来做一些对安全性不在意的数据收集，如意见反馈、问卷调查。如果你拿它来收集用户的银行卡号，用于给用户打款，就存在显著的安全隐患了。

3. 联系人管理

你可能注意到了，在金数据的导航栏上，还有一个联系人模块。这里会收集所有通过表单获取到的联系人身份和信息，看到每一个人填写过的所有历史表单内容。

金数据的联系人模块，点击查看个人详情，可以看到他填写过的所有历史表单

同时，你还可以根据不同表单来筛选联系人信息，并给他们打上标签。例如，你作为产品经理，每次版本升级后都会给用户发放新功能调研问卷，你可以把填过这些问卷的人，都

打上「核心用户」的标签。然后，通过金数据的群发邮件/短信的功能，邀请这些人加入核心用户微信群，方便进一步的运营。

可以批量选中一批人，或按筛选条件、标签等，给一批用户群发短信/邮件

美中不足的是，金数据只能按用户填过什么表单来打标签，而暂时还不能按用户填写的内容来打标签。比如，作为产品经理，我想把多份表单内，填写过喜欢我们的产品，并活跃在多个社交平台的用户筛选出来，统一打上「社交传播者」的联系人标签，然后群发给这批用户，引导他们分享我们的产品。

要做到这一点，现在的联系人功能并不能做到，只能曲线实现：到每一份表单中，按照「第二章：无需开发，搭建一套业务系统」中「通知能力：按照用户选型群发」的办法，按用户填写的表单内容，过滤出选型用户后再群发。这样的办法，稍显麻烦之外，还存在的问题是，如果同一个用户在多个表单中多次被过滤出来，会重复收到多次群发短信/邮件。

现在，你已经了解了想要收集高质量数据的基本思路，就是尽可能地要满足三个标准：真实性、结构化和标准化、和人关联。要做到这些，有些可以借助金数据的设置和选项，帮你更轻松地实现目标，例如防止重复提交、获取微信用户信息等；有些则需要你在题目的设置上留有一些技巧，如在问题设置时交叉验证，以及引导用户按照统一的标准输入等。

以上，就是如何借助金数据来收集高质量的数据。如果你还有什么补充和建议，也欢迎在评论中和大家分享。